BRAC University

Sentiment Analysis for Bengali Newspaper Headlines

By

Mohammad Samman Hossain

ID: 13301040

Israt Jahan Jui

ID: 13301120

Afia Zahin Suzana

ID: 13101289

Supervised by

Dr. Md. Haider Ali

Professor

Department of Computer Science and Engineering

BRAC University

Submission Date:

Declaration

This is to certify that the research work titled “Sentiment Analysis for Bengali Newspaper Headlines” is submitted by Afia Zahin Suzana, Israt Jahan Jui and Mohammad Samman Hossain to the Department of Computer Science & Engineering, BRAC University in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering. We hereby declare that this thesis is based on results found from our own work. Materials of work found by other researcher are mentioned by reference. This Thesis, neither in whole or in part, has been previously submitted for any degree. We carried our research under the supervision of Professor Md. Haider Ali.

Signature of Supervisor

_____

Dr. Md. Haider Ali

Signature of Author

_____

Mohammad Samman Hossain

Signature of Author

_____

Israt Jahan Jui

Signature of Author

_____

Afia Zahin Suzana

Abstract

In this project, we propose a system that assigns scores indicating positive or negative to newspaper headlines. Many works have been done on sentiment analysis, document clustering for newspaper headlines in English. We are going to do the same for Bengali language. News headlines from one Bengali newspaper [1] is used for the purpose of the project. As there is no dataset of headlines for this newspaper, we are using web crawler to get necessary headlines to make a dataset to use for this project. For our proposed system, a number of classifiers will be used such as Support Vector Machine [2], Logistic Regression [17] etc. Through the experiments, our aim is to establish a system, which can identify positive and negative news accurately.

Keywords: sentiment analysis, SVM, Boosted Tree, Logistic, Bengali, classifiers

Acknowledgement

Our deep thanks to our thesis supervisor Dr. Md. Haider Ali for his guidance and continuous support throughout the work.

We are extremely thankful to our family, friends for their support and encouragement.

Finally, we thank BRAC University for giving us the opportunity to complete our BSc. Engineering degree in Computer Science and Engineering.

Table of Contents

List of Figures

List of Tables

Abbreviations

SVM – Support Vector Machine

CSV – Comma Separated Values

1. Introduction

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space [4]. For our work, we are focusing on newspaper headline. News can be positive, negative, or neutral. Although full understanding of natural text language is well beyond capabilities of machines, statistical analysis of relatively simple statement can provide a surprisingly meaningful categorization of positive and negative sentiment. Analyzed data then can be used in many sectors. It can be used to predict financial state of a country and other predictions.

Sentiment Analysis of English newspaper headline had been done on many occasions. Research in such an area was difficult in the past, due to the scarcity of large manually annotated corpora [5]. This problem has been solved because of the recent increase in the number of news websites.

News Articles can be collected daily by manually copying the news or using RSS feeds that are mostly available in every news website like CNN, BBC News etc. By serving as data sources, these websites facilitate the research on identifying the polarity of the news articles that can be helpful to readers as this can be beneficial to them. For our project, we have chosen to analyze the sentiment of Bengali newspaper headlines. Bengali is one of the highest spoken language, ranked seventh in the

world, but surprisingly very few works had been done with Bengali language on sentiment analysis [6].

In this project, we are going to extract the sentiment conveyed by one Bengali newspaper [1] and will identify the polarity of the headlines as positive or negative. Through sentiment analysis, we are going to classify given headlines into one of three categories – positive, negative or neutral.

2. Literature Review

Our work is mainly inspired by Kaur and Chopra's work [5]. They classified newspaper articles into three polarity. Positive, Negative and Neutral. In our work, we are going to implement SVM [2], Logistic Regression [17] and Boosted Tree classifiers.

In [5], they used Naive Bayes classifier, and it is done on English. Actually, most of the work done on sentiment analysis are done on English. Work on other natural languages is increasing including Japanese [7] [8] [9] [10], Chinese [11] [12] and other languages. Works that are done in Bengali are done regarding blog's [13] [14].

We choose Logistic classifier because in comparison with Naïve bayes classifier, Logistic classifier always preferred. Logistic classifier [17] has less asymptotic error [17] than Naïve Bayes classifier.

In [15], sentiment analysis was done on news and on blogs. Their system consisted of sentiment identification, aggregation and scoring phase. Wanner et. Al [16] presented a visual tool for large-scale comparative sentiment analysis of news articles.

In [22], Logistic regression makes minimal measure of assumption on the dataset, it measures the connection between categorical binary dependent variable and the independent variables. However, it appraises the probabilities of the categorical variable using the logistic function.

3. Methodology

3.1 Data Collection

For our work, we used the headlines from prothom-alo.com. No dataset of headlines existed at that time, so we needed to create our own dataset. We used web crawler to get the source html script of the website of Prothom-alo. We pulled html script programmatically of all the pages that were published in that newspaper from January 2016 to May 2016; also, we used regular expression to extract the headlines from the source html script of the site. We saved the headlines in a text file and every month had their own text file.

3.2 Data Formatting

We saved the headline along with their respective dates in our dataset. We separated the headlines from the dates by putting "||" in the middle of them.

3.3 Data Tagging

To train our models we needed manually tagged data. To make the process of data tagging easier we wrote a code that automated the whole data tagging process. It read the headlines one by one and tagged those headlines positive, negative or neutral. After that, we took all the month-by-month text files and comma separated the headlines, date, and tag. This process was also automated by writing a piece of code. We then saved all the data in one csv file, and this csv file worked as our main dataset.

3.4 Classifier Selection

There are many machine-learning classifiers for text classification. After researching, we decided to use SVM, Boosted Tree and Logistic Regression in our

project. SVM has been found providing better accuracy in the case of classifying text. As SVM and Logistic Regression are binary classifiers, they are better suited in classifying polarity of a sentence. Since our work is to differentiate between positive and negative which is like binary classification and SVM, Logistic Regression works better for it. Boosted Tree classifier do not expect linear features or even features that interact linearly, which is different from Support Vector Machine and Logistic Regression. We wanted to see how it performs against Logistic Regression and Support Vector Machine.

3.5 Work Flow

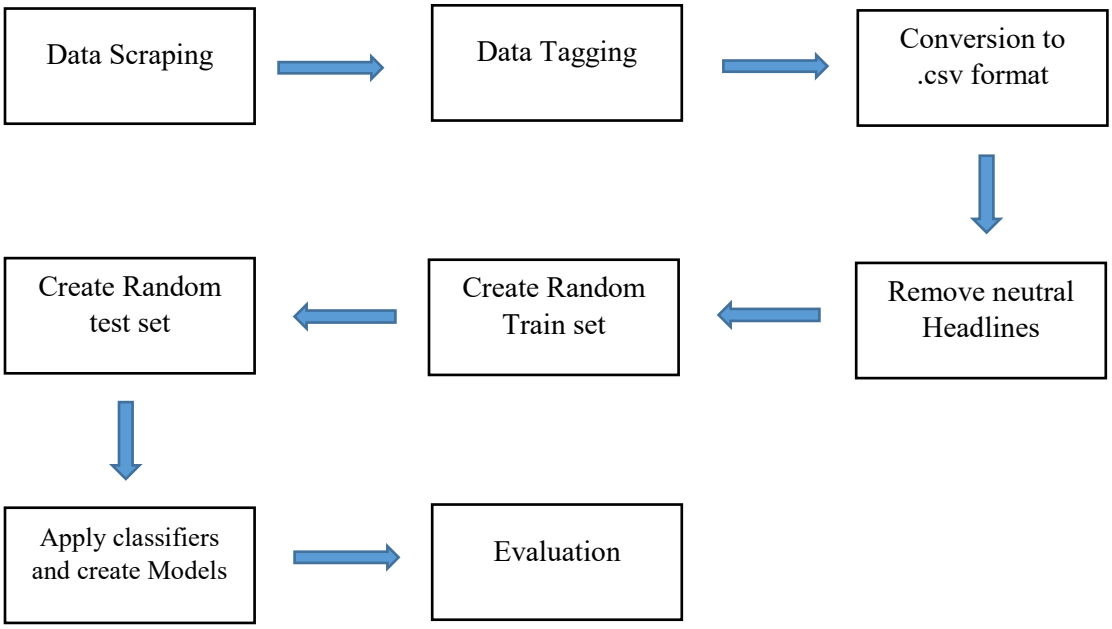

Fig 3.1: Work Flow

4. Machine Learning Algorithms

4.1 Support Vector Machine (SVM)

SVM (Support Vector Machine) is a machine-learning algorithm. SVM is a supervised learning model, which analyzes data for classification and regression analysis. SVM build a model using training algorithm that assigns new examples into one or two categories. SVM divided categories as wide as possible by creating a gap. New applications categories gap mapped into that same space or gap on which side the application fall on. The problem is when data are not properly labelled supervised learning is not possible. Then we have to follow an unsupervised learning approach to analyze, by which we can divide the data into separate groups. It follows a clustering approach that is called support vector clustering [18] and is often used in industrial applications either when data is not labelled or when only some data is labelled.

Example

Outlier: An outlier is an observation point that is distant from other observations. An observation that is well outside of the expected range of values in a study or experiment, and which is often discarded from the dataset.

Hyper plane: In geometry, a hyper plane is a subspace of one dimension less than its ambient space. If a space is 3-dimensional then its hyper planes are the 2-dimensional planes, while if the space is 2-dimensional, its hyper planes are the 1-dimensional lines. This notion can be used in any general space in which the concept of the dimension of a subspace is defined. Suppose, we have three hyper-planes (A, B and C). Now, we need to identify the right hyper-plane to classify star and circle. We need to remember a thumb rule to identify the right hyper-plane.

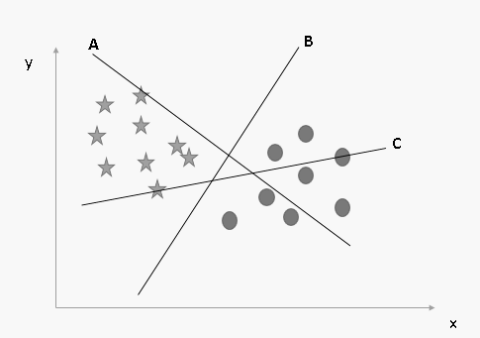

Fig 4.1: Hyper plane in scatter

Now, here we have three hyper plane all are in scatter possible ways. Now, how can we identify the right hyper-plane from these?

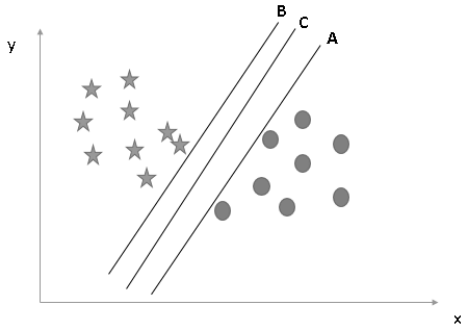

Fig 4.2: Creating Hyper Plan margin

Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as Margin.

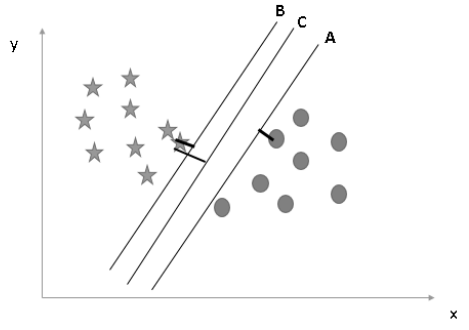

Fig 4.3: Finding right Hyper Plane

Above, you can see that the margin for hyper-plane C is higher than A and B. Hence, we name the right hyper-plane as C. We have selected the hyper-plane with higher

margin is robustness [18], because if we select a hyper plane having low margin then there is a high chance is that the margin will be miss-classified.

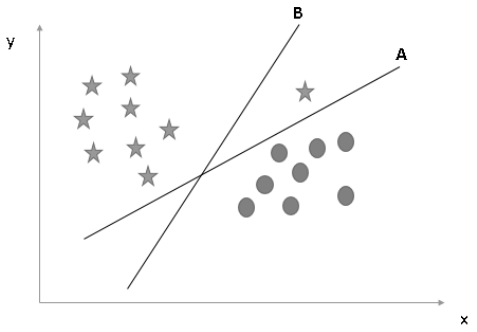

Fig 4.4: Correct Hyper Plane

Here, one might select B as the higher margin, but SVM selects the hyper plane, which classifies the classes accurately prior to maximizing margin. B has a classification error and A has classified all correctly. Therefore, right hyper plane is A.

Below we cannot segregate two classes with a straight line, as one of the star is in the territory of other.

Fig 4.5: Hyper Plane with Outliers

SVM has a feature to ignore outliers and find the hyper plane that has maximum margin. Therefore, it can be said that, SVM is robust to outliers.

In the scenario below, a linear hyper plane between two classes is not possible.

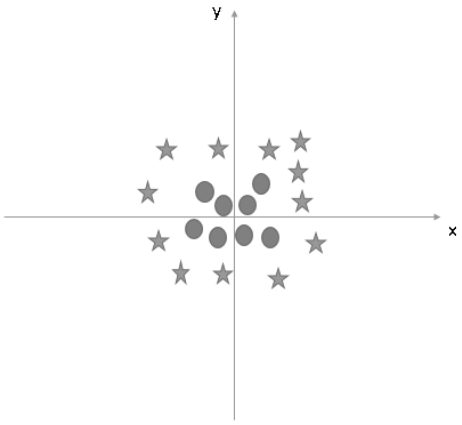

Fig 4.6: Hyper plane with maximum margine

This problem also can be solved by SVM. SVM introduces an additional feature z=x^2+y^2. If we plot the data points on axis x and z:

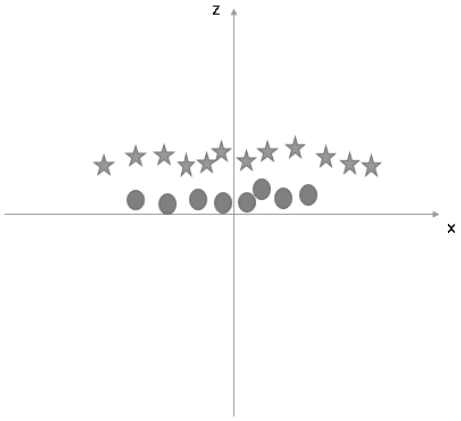

Fig 4.7: Solving with SVM

SVM Applications

SVM is a device for content classification, which can lessen the requirement for marked preparing cases in both standard inductive and tranductive settings. Picture classification, which is a piece of picture preparing, can likewise be performed utilizing support vector machine. In the examination and result investigation SVM accomplish higher precision that other customary plans after only at least three round of input. To discover the picture classification SVM takes after same customary approach as ordinary content examination. The SVM algorithm has been additionally broadly utilized as a part of the natural and different sciences to discover results. SVM classification have been utilized and it offers up to 90% of the compound effectively. Support vector machine weights have additionally been utilized to consider SVM models previously. Psthoc interpretation of support vector machine models has been utilized as a part of request to recognize components is a generally new territory of research with unique importance in the organic sciences [19].

4.2 Logistic Regression

Every machine-learning algorithm works best under a given set of conditions. Making sure that the algorithm fits the requirements ensures superior performance. Any algorithm in any condition cannot be used. For example: We can not use linear regression on a categorical dependent variable. Instead, in such condition SVM, Logistic Regression should be used. Logistic Regression is a classification algorithm. It is used to predict a binary outcome, given a set of independent variables. To represent binary or categorical outcome, dummy variables are used. Logistic Regression is like a special case of linear regression when outcome cariable is categorical.

Given a set [20] of features x_i, and a label $y_i \in \{0,1\}$, logistic regression interprets the probability that the label is in one class as a logistic function of a linear combination of the features:

$$fi(\theta) = p(y_i = 1|x) = \frac{1}{1 + \exp(-\theta^T x)}$$

Analogous to linear regression, an intercept term is added by appending a column of 1's to the features and L1 and L2 regularizers are supported. The composite objective being optimized for is the following:

$$\min(\theta) \sum_{i=1}^{n} f_i(\theta) + \lambda_1||_1 + \lambda_2||\theta||_2^2$$

where λ1 is the L1_penalty and λ2 is the L2_penalty.

Multiclass Classification

Multiclass classification is the problem of classifying instances into one of many (i.e more than two) possible instances. As an example, binary classification can be used to train a classifier that can distinguish between two classes, say "cat" or "dog", while multiclass classification can be used to train a finite set of labels at the same time, say "cat", "dog", "rat", and "cow".

Top-k predictions

Multiclass classification provides the top-k class predictions for each class. The predictions are either margins, probabilities, or a rank for the predicted class for each example. In the following example, we provide the top 5 predictions, ordered by class probability, for each data point in the test set.

4.3 Boosted Tree

The Gradient Boosted Regression Trees (GBRT) model (also called Gradient Boosted Machine or GBM) is one of the most effective machine learning models for predictive analytics [21].

Background

The Boosted Trees Model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \cdots$$

where the final classifier g is the sum of simple base classifiers fi. For boosted trees model, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called model ensembling.

Unlike Random Forest which constructs all the base classifier independently, each using a subsample of data, GBRT uses a particular model ensembling technique called gradient boosting.

The name of Gradient Boosting comes from its connection to the Gradient Descent in numerical optimization. Suppose you want to optimize a function f(x), assuming f is differentiable, gradient descent works by iteratively find

$$x_{t+1} = x_t - \eta \frac{\partial f}{\partial x}|_{x=x_t}|$$

where η is called the step size.

Similarly, if we let gt(x)=∑t−1i=0fi(x) be the classifier trained at iteration t, and L(yi,g(xi)) be the empirical loss function, at each iteration we will move gt towards the negative gradient direction −∂L∂g||g=gt by η amount. Hence, ft is chosen to be

$$f_t = \arg\min(f) \sum_{i=1}^{N} \left[\frac{\partial L\left(y_i, g(x_i)\right)}{\partial g(x_i)} \Big|_{g=g_t} - f(x_i) \right]^2$$

and the algorithm sets $g_{t+1} = g_t + \eta f_t$

Use of Boosted Tree

Different kinds of models have different advantages. The boosted trees model is very good at handling tabular data with numerical features, or categorical features with fewer than hundreds of categories. Unlike linear models, the boosted trees model are able to capture non-linear interaction between the features and the target.

One important note is that tree based models are not designed to work with very sparse features. When dealing with sparse input data (e.g. categorical features with large dimension), we can either pre-process the sparse features to generate numerical statistics, or switch to a linear model, which is better suited for such scenarios.

5. Experiment & Result Analysis

5.1 Experiment

We gathered and tagged 15325 headlines for our dataset. This number consists of all the negative, positive and neutral headlines. Before using the classifiers, we had to work on the dataset, so we can use it for creating the model. This was the head of the dataset before anything was changed.

Sentiment	Headline	Date
Negative	জঙ্গি কর্মকাণ্ডে বিশ্ববিদ্যালয়ের ৭২ ছাত্র–শিক্ষক	2016-10-01
Negative	বিভীষিকার সেই রাত	2016-10-01
Neutral	মূল হোতাদের বিচার চায় নিহত জঙ্গিদের পরিবার	2016-10-01
Positive	ইইউ প্রতি বর্গ কিলোমিটারেই নিরাপত্তা চায়	2016-10-01
Negative	এক রাতেই ভেসে গেল রেণু বেওয়ার ঘর–সংসার	2016-10-01
Negative	জঙ্গিবাদবিরোধী মানববন্ধনে সংঘর্ষে জড়াল ছাত্রলীগ	2016-10-01
Neutral	দিনে তিন হাজার বাংলাদেশি ভারতে যায়: স্বরাষ্ট্রমন্ত্রী	2016-10-01
Neutral		2016-10-02
Negative	ছাত্রলীগের কোন্দলে এবার কুমিল্লায় খুন	2016-10-02
Negative	৮০০ শিক্ষাপ্রতিষ্ঠানে পাঠদান বন্ধ	2016-10-02

Table 5.1 Head of the Dataset

First, we took out all of the headlines that were tagged neutral. Reason behind this is, we were working to get a model that can predict the positive and negative headlines, not the neutral headlines. After changing the dataset, we were left with only 3120 headlines, which are positive or negative. The head of our csv file was like this.

Sentiment	Headline	Date
Negative	জঙ্গি কর্মকাণ্ডে বিশ্ববিদ্যালয়ের ৭২ ছাত্র–শিক্ষক	2016-10-01
Negative	বিভীষিকার সেই রাত	2016-10-01
Positive	ইইউ প্রতি বর্গ কিলোমিটারেই নিরাপত্তা চায়	2016-10-01
Negative	এক রাতেই ভেসে গেল রেণু বেওয়ার ঘর–সংসার	2016-10-01
Negative	জঙ্গিবাদবিরোধী মানববন্ধনে সংঘর্ষে জড়াল ছাত্রলীগ	2016-10-01
Negative	ছাত্রলীগের কোন্দলে এবার কুমিল্লায় খুন	2016-10-02
Negative	৮০০ শিক্ষাপ্রতিষ্ঠানে পাঠদান বন্ধ	2016-10-02
Negative	জাল দলিলে জমি কেনাবেচা	2016-10-03
Negative	তিন বন্ধুর অস্ত্র প্রশিক্ষণ মিরপুরে	2016-10-04
Negative	গোড়াপত্তন আফগান মুজাহিদদের হাতে	2016-10-04

Table 5.2: Head of the dataset without neutral tagged data

Number of headlines that were tagged negative was much more than the number of data that was tagged positive. Only 749 out of 3120 headlines was positive, all the other was negative.

Sentiment	Headline	Date
Positive	ইইউ প্রতি বর্গ কিলোমিটারেই নিরাপত্তা চায়	2016-10-01
Positive	সবজির বীজে স্বাবলম্বী হাসেনা বেগম	2016-10-06
Positive	বাংলাদেশেরও ‘অন্য রকম’ অলিম্পিক শুরু আজ	2016-10-05
Positive	সুন্দরবন ঘিরে ১৫০ শিল্প প্রকল্প!	2016-10-09
Positive	গুলশানের হামলা মামলায় গ্রেপ্তার হাসনাত করিম	2016-10-14
Positive	চার বছর পর হিমাদ্রীর মায়ের ‘স্বস্তি’	2016-10-15
Positive	বিসিএসে এবার মেধায় নিয়োগ ৬৭ শতাংশ	2016-10-18
Positive	বাংলাদেশেরও গৌরব রিতা	2016-10-22
Positive	দীপন হত্যার মূল আসামি গ্রেপ্তার	2016-10-25
Positive	বন্যাকবলিত এলাকায় যাচ্ছেন আ.লীগের কেন্দ্রীয় নেতারা	2016-10-01

Table 5.3: Head of the dataset with only positive headlines

Sentiment	Headline	Date
Negative	জঙ্গি কর্মকাণ্ডে বিশ্ববিদ্যালয়ের ৭২ ছাত্র-শিক্ষক	2016-10-01
Negative	বিভীষিকার সেই রাত	2016-10-01
Negative	এক রাতেই ভেসে গেল রেণু বেওয়ার ঘর–সংসার	2016-10-01
Negative	জঙ্গিবাদবিরোধী মানববন্ধনে সংঘর্ষে জড়াল ছাত্রলীগ	2016-10-01
Negative	ছাত্রলীগের কোন্দলে এবার কুমিল্লায় খুন	2016-10-02
Negative	৮০০ শিক্ষাপ্রতিষ্ঠানে পাঠদান বন্ধ	2016-10-02
Negative	জাল দলিলে জমি কেনাবেচা	2016-10-03
Negative	তিন বন্ধুর অস্ত্র প্রশিক্ষণ মিরপুরে	2016-10-04
Negative	গোড়াপত্তন আফগান মুজাহিদদের হাতে	2016-10-04
Negative	পুলিশের ধাওয়ায় পানিতে ডুবে ব্যবসায়ীর মৃত্যু পিটুনিতে পুলিশ নিহত	2016-10-04

Table 5.4: Head of the dataset with only negative headlines

To create a good model that can accurately predict headlines, we needed to make the number of positive and negative headlines near to equal. We chose 31% negative headlines and we chose it randomly, and we got 787 negative headlines. This dataset with 749 positive and 787 negative headlines is the main dataset with which we are going to create our model and find accuracy.

We split this data in two parts. Training data and test data Headlines of every part are selected randomly. Training consists of 90% of the headlines, which are going to be used to train the model that we are going to create. Other 10% are the test data, which we will use to test the accuracy of our created model.

We added a new column in our dataset name “Words”, this would be the feature vector for the models that we create.

5.2 Experiment with Logistic

With the dataset we got from experiment part, we trained our model with logistic classifier. We selected the words as our feature vector, and limited max iteration to 200.

Number of examples : 1399
Number of classes : 2
Number of feature columns : 1
Number of unpacked features : 3308
Number of coefficients : 3309

The training accuracy for Logistic Classifier is shown in the table below:

Iteration	Passes	Elapsed Time	Training Accuracy
1	3	1.053230	0.989993
2	5	1.093740	0.995711
3	6	1.094739	0.997141
4	7	1.097743	0.998570
5	8	1.102745	0.998570
6	9	1.104747	0.999285
11	14	1.117756	0.999285

Table 5.5: Training Accuracy of Logistic Classifier

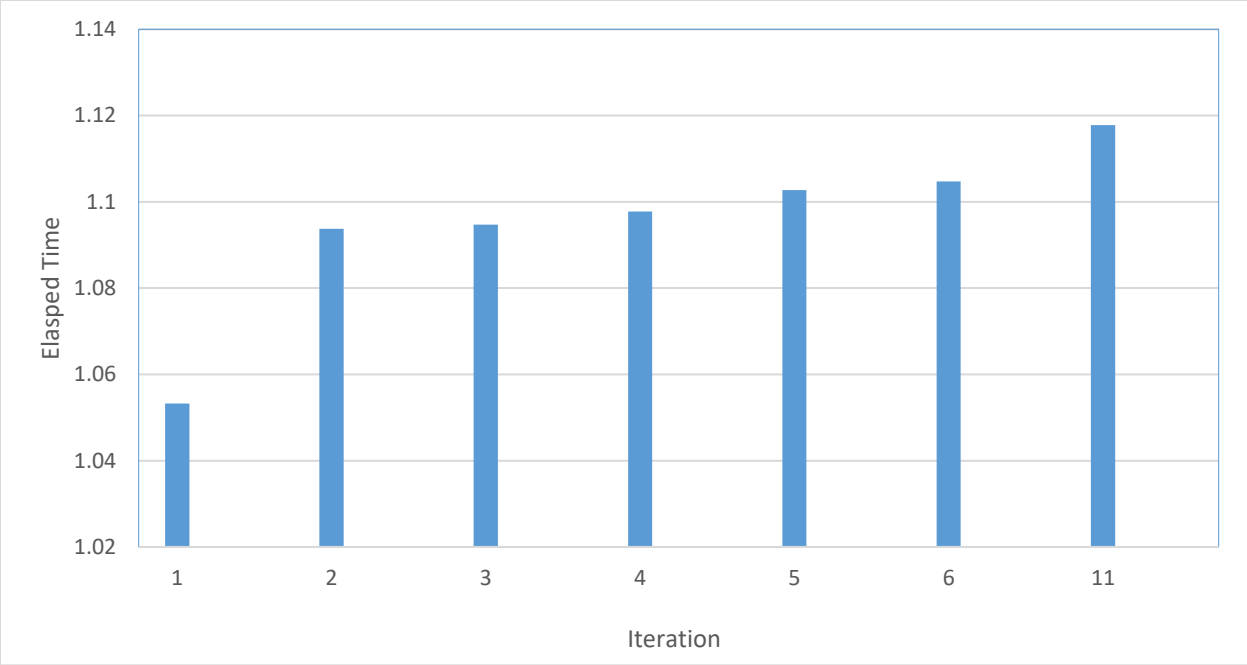

Fig 5.1: Logistic Classifier: Iteration vs Elapsed Time

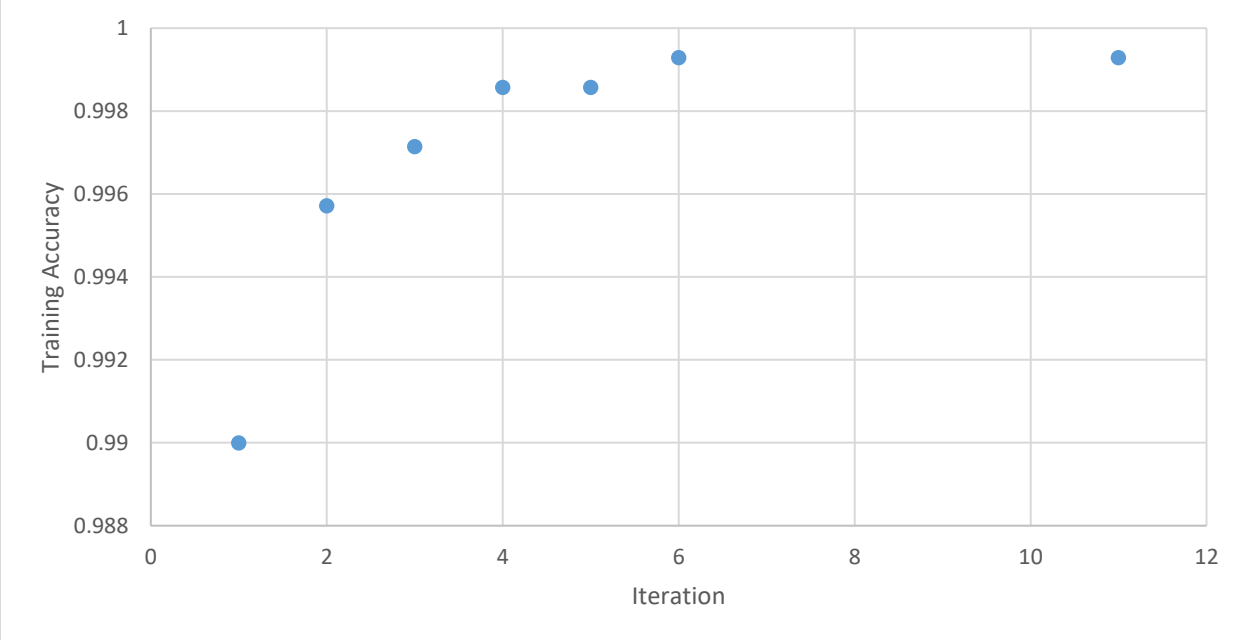

Fig 5.2: Logistic Classifier: Iteration vs Training Accuracy

Although max iteration was set to 200, we reached optimum solution in 11 iterations. Test accuracy for Logistic Classifier was **0.759124087591240**

5.3 Experiment with SVM

We used the same dataset that we used with logistic classifier to experiment with Support vector machine. Same dataset was used so that we can compare all the classifiers that would be used. With words as feature vector, we set the max iteration for this classifier to 2000

Number of examples : 1399

Number of classes : 2

Number of feature columns : 1

Number of unpacked features : 3308

Number of coefficients : 3309

The training accuracy is for svm classifier shown in the table below:

Iteration	Passes	Elapsed Time	Training Accuracy
1	3	0.003001	0.989993
2	5	0.007004	0.996426
3	6	0.010006	0.997856
4	7	0.012008	0.997856
5	8	0.014008	0.998570
6	9	0.016010	0.998570
11	14	0.026522	0.989993
51	88	0.140645	0.999285
101	205	0.297235	0.999285

Table 5.6: Training Accuracy of SVM Classifier

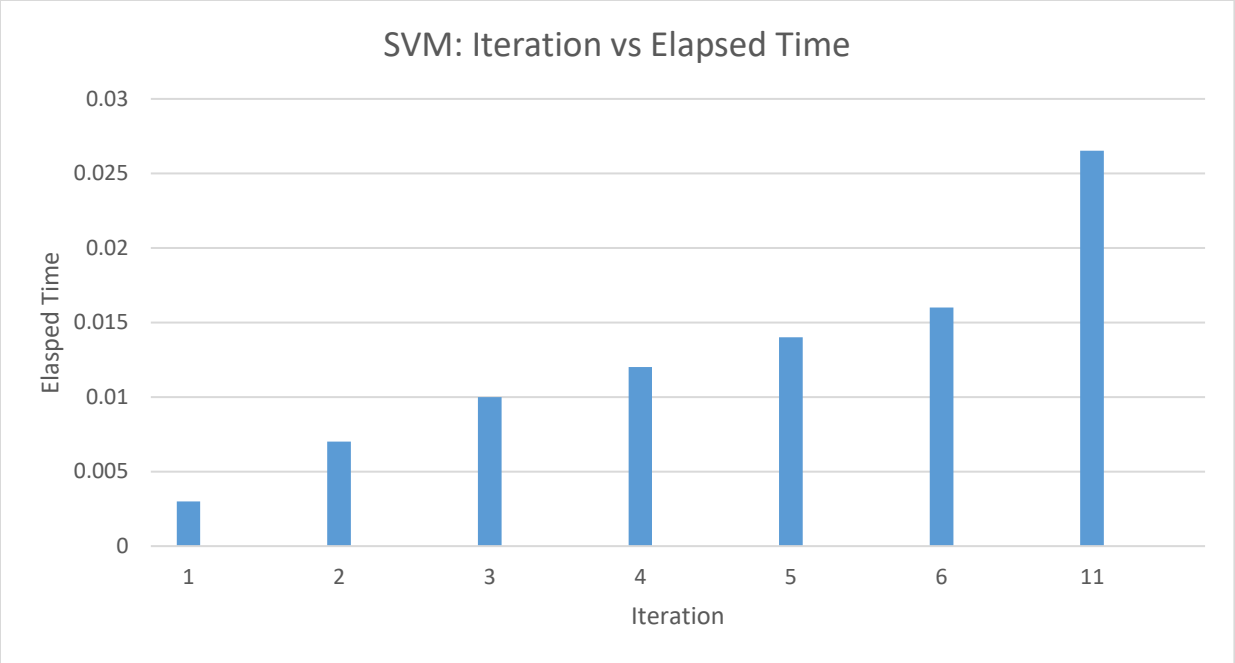

Fig 5.3: SVM: Iteration vs Elapsed Time

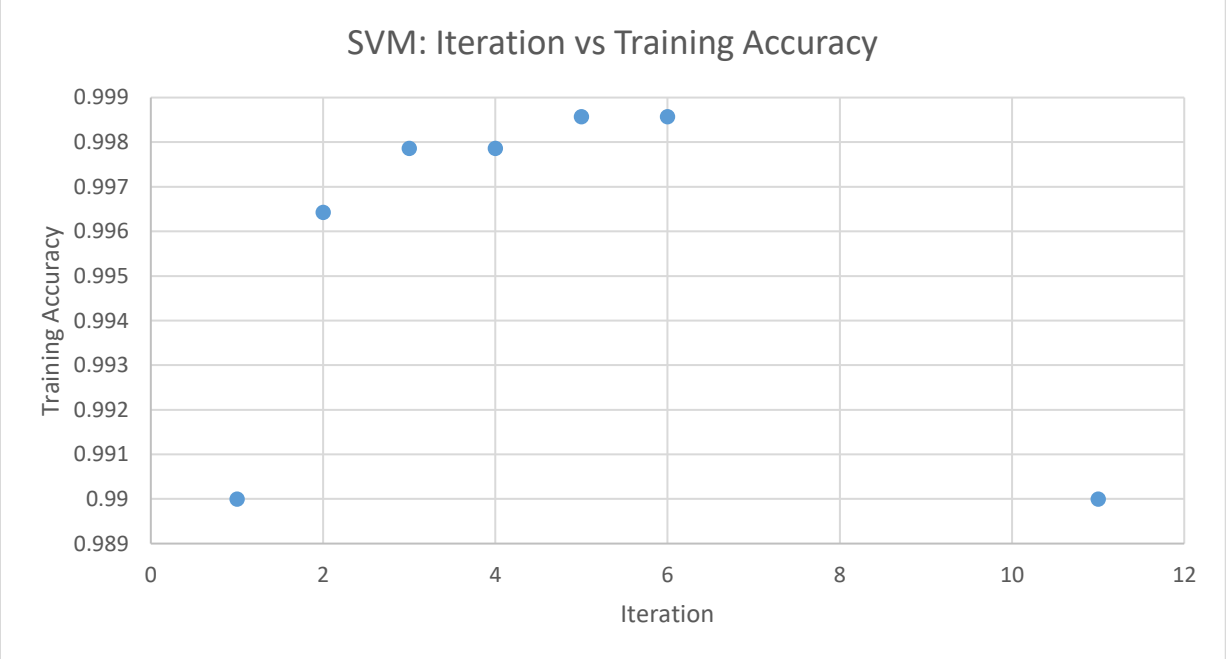

Fig 5.4: SVM: Iteration vs Training Accuracy

Although max iteration was set to 2000, we reached optimum solution in 101 iterations. Test accuracy for SVM Classifier was **0.7956204379562044**

5.4 Experiment with Boosted Tree

We set the max iteration of boosted tree classifier to 1000

Number of examples : 1399
Number of classes : 2
Number of feature columns : 1
Number of unpacked features : 3308

The training accuracy is for boosted tree classifier shown in the table below:

Iteration	Elapsed Time	Training Accuracy	Training Log Loss
1	0.023016	0.645461	0.638970
2	0.042033	0.641887	0.607912
3	0.046535	0.652609	0.586641
4	0.049537	0.655468	0.570033
5	0.054043	0.669049	0.555089
6	0.058047	0.676197	0.545534
11	0.082064	0.725518	0.506033
51	0.204158	0.852752	0.379590
101	0.323789	0.917084	0.312295
250	0.651848	0.960686	0.216643
500	1.308713	0.998570	0.145822
501	1.313715	0.998570	0.145652
750	1.905286	0.998570	0.107537
1000	2.450225	0.998570	0.083154

Table 5.7: Training Accuracy of Boosted Tree classifier

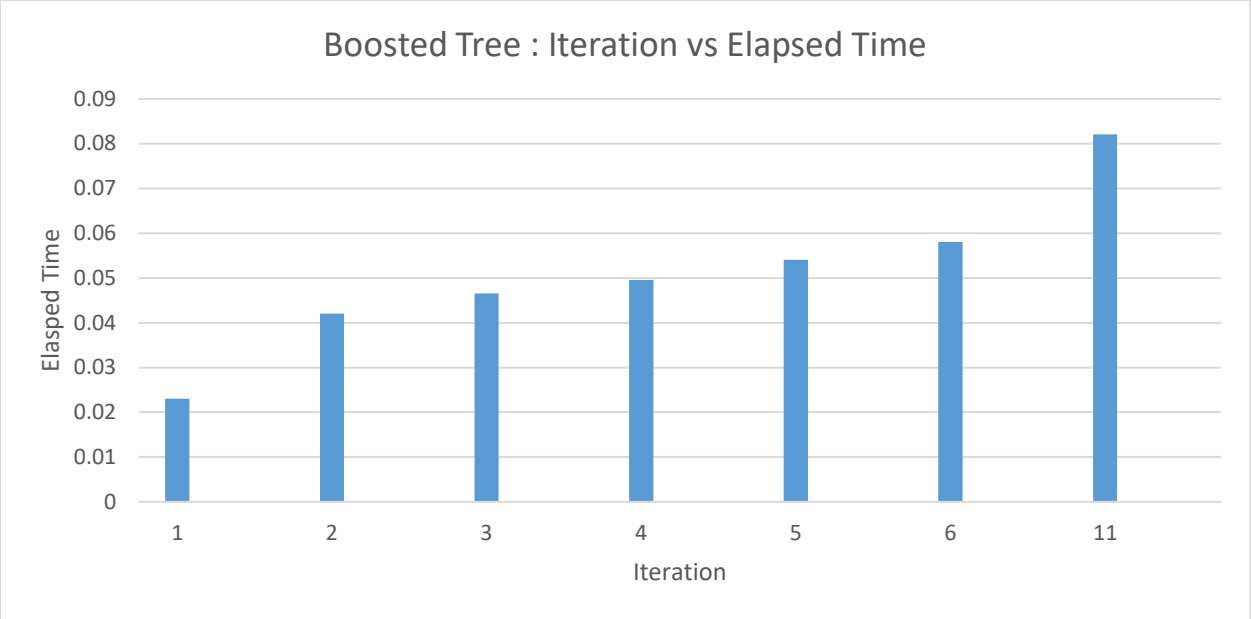

Fig 5.5: Boosted Tree: Iteration vs Elapsed Time

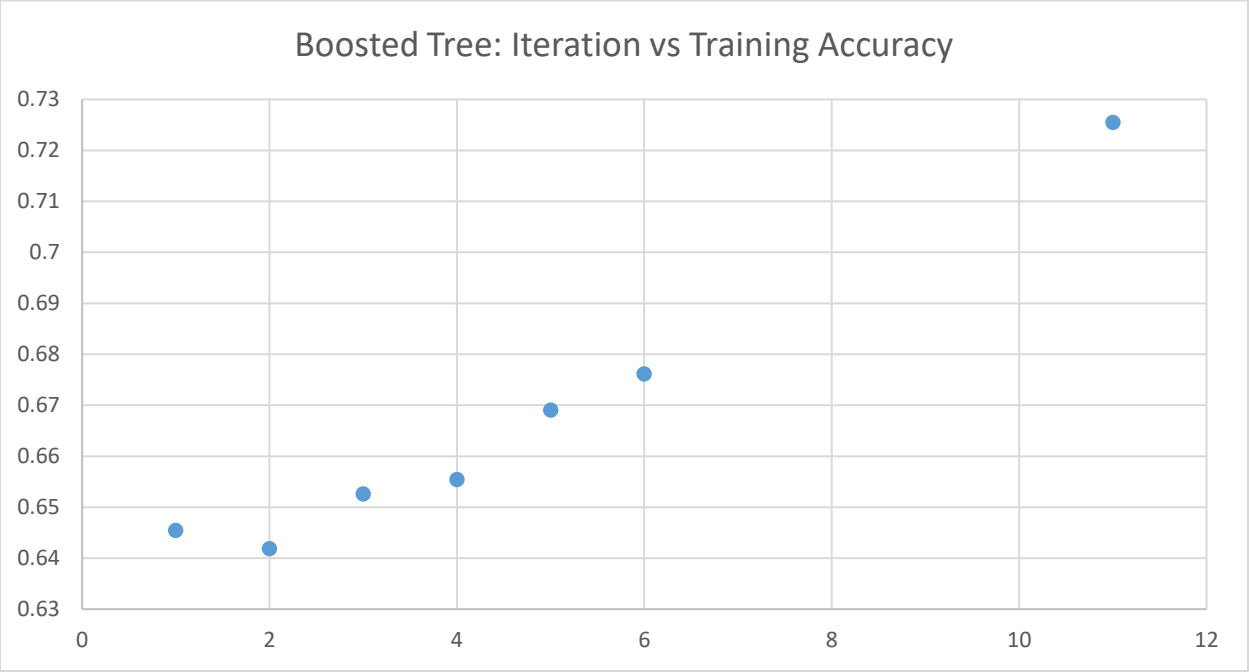

Fig 5.6: Boosted Tree: Iteration vs Training Accuracy

Test accuracy for Boosted Tree Classifier was **0.7664233576642335**

5.5 Comparative Analysis

We are comparing the classifiers with their F1 score, Precision, Recall, Log loss, Final Training (After 11 iteration) and Test accuracy.

Precision and Recall

Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

F1 Score

The F1 score is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0

	Logistic	SVM	Boosted Tree
F1 Score	0.755556	0.791044	0.783784
Precision	0.822581	0.868852	0.773333
Recall	0.698630	0.726027	0.794521
Training Accuracy	0.999285	0.989993	0.725518
Test Accuracy	0.759124	0.795620	0.766423

Table 5.8: Comparison

As we can see from the Table 5.8, SVM has the highest F1 score, which indicates svm classifiers test was most accurate for this dataset. SVM has the highest precision score too, and it indicates that the ratio of correct prediction vs search results of svm is higher than other two classifiers. While boosted tree has the highest recall value. It indicates that the ratio of correct prediction vs whole dataset is highest in boosted tree.

We compared training accuracy of the classifiers after 11 iterations are completed. SVM and Logistic classifiers accuracy was nearly equal with accuracy greater than 0.98. While boosted tree had training accuracy of 0.76.

Test accuracy of all the classifiers are:

Logistic Classifier: **0.759124087591240**

SVM Classifier: **0.7956204379562044**

Boosted Tree Classifier: **0.7664233576642335**

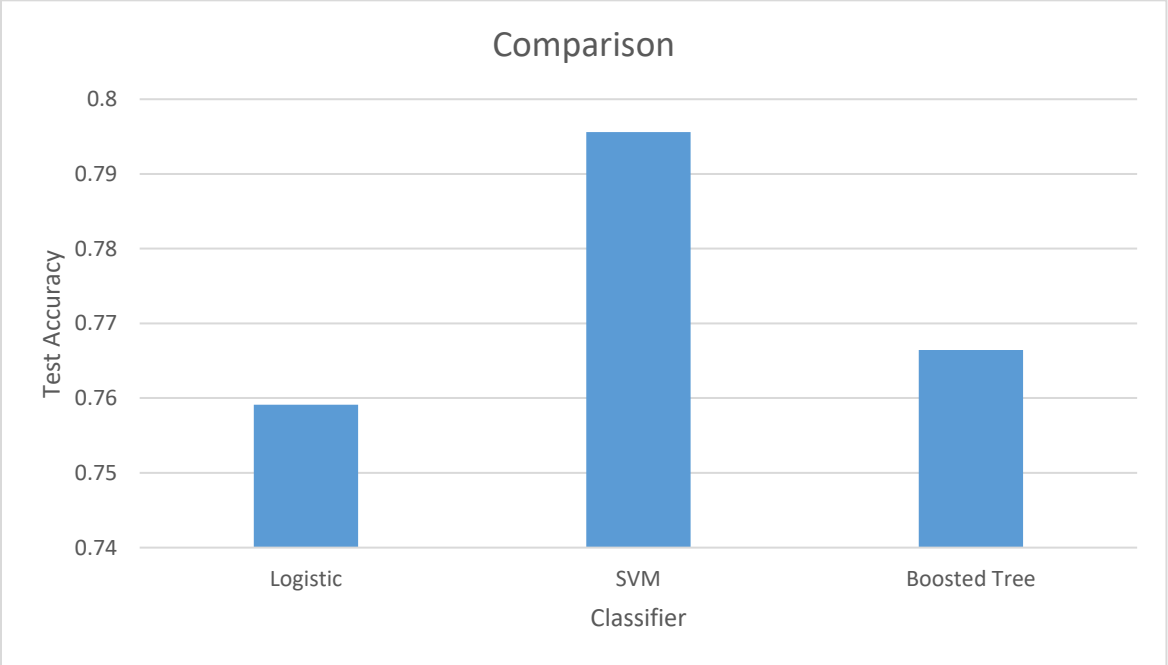

Fig 5.7: Comparison of test accuracy

As we can see, test accuracy of the classifiers are from 0.75 to 0.80, where SVM has the highest accuracy. Logistic and boosted tree classifier had similar accuracy. Difference between SVM and other two classifier that we used is that, SVM is a kernel based classifier. Therefore, for our project kernel based classifier has the best accuracy.

6. Conclusion

In our thesis work, we tried to create a model which can predict if a newspaper headline is positive or negative. We created the model with several months headline from a Bengali daily newspaper Prothom-alo. There was no dataset with the headlines from that newspaper. Therefore, we had to collect the data by scraping their website. After collecting the data, we tagged them manually. We only collected 15000 headlines but had to delete 13000 headlines because they were neutral headlines. Our dataset had 2000 headlines then. We used a dataset of 1400 headlines cause number of positive and negative headlines had to be nearly equal. Therefore, we deleted 700 more headlines. For sentiment analysis, this was a small dataset, and because of this, our model became over fitted and we got higher than normal training accuracy. There is scope to create a model that would predict headlines more accurately, but for that our dataset would have to consist a huge amount of data that we have to scrap and then tag them manually. We plan to work on getting years' worth of headlines and create a dataset that anyone can use for their research. Moreover, we will also see, if number of positive or negative news has any impact on daily life, stock market etc.

7. References

[1] http://www.prothom-alo.com/

[2] Mullen T. and Collier N., “Sentiment analysis using support vector machines with diverse information sources”, National Institute of Informatics (NII), Japan

[3] Gamallo P. and Garcia M., “Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets” , Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 171–175,Dublin, Ireland, August 23-24 2014.

[4] Liu B., “Sentiment Analysis and Opinion Mining”, Morgan & Claypool Publishers, May 2012.

[5] Kaur H. and Chopra V., “Sentiment Analysis of News Headlines using Naive Bayes Classifier”, International Journal for Science, Management and Technology

[6] Chowdhury S. and Chowdhury W., “Sentiment Analysis for Bangla Microblog Posts”, BRAC University, Dhaka, Bangladesh

[7] Kanayama H. and Nasukawa T., “Fully automatic lexicon expansion for domain-oriented sentiment analysis”. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 2006.

[8] Kobayashi N., Inui K., Tateishi K., and Fukushima T., “Collecting evaluative expressions for opinion extraction”. In Proceedings of IJCNLP 2004, pages 596–605, 2004.

[9] Suzuki Y., Takamura H., and Okumura M., “Application of semi-supervised learning to evaluative expression classification”. In Proceedings of the 7th International Conference on Intelligent Text Processing and Computational Linguistics, 2006.

[10] Takamura H., Inui T., and Okumura M., “Latent variable models for semantic orientations of phrases”. In Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics, 2006.

[11] Hu Y., Duan J., Chen X., Pei B., and Lu R., “A new method for sentiment classification in text retrieval”. In IJCNLP, pages 1–9, 2005.

[12] Zagibalov T. and Carroll J., “Automatic seed word selection for unsupervised sentiment classification of chinese text”. In Proceedings of the Conference on Computational Linguistics, 2008.

[13] Das A.and Bandyopadhyay S., (2009a). “Subjectivity Detection in English and Bengali: A CRF-based Approach.”, In Proceeding of ICON 2009, December 14th-17th, 2009, Hyderabad.

[14] Das D., Bandyopadhyay S., “Labeling emotion in Bengali blog corpus–a fine grained tagging at sentence level”, In Proceedings of the 8th Workshop on Asian Language Resources, pages 47–55, Beijing, China, August 2010.

[15] Srinivasaiah M., Godbole N. and Skiena S., “Large Scale Sentiment Analysis for News and Blogs”, Stony Brook University, Stony Brook, NY 11794-4400, USA

[16] Wanner F., Rohrdantz C., Mansmann F., Stoffel A., Oelke D., Kristajic M., Keom D.A. Luo D., Yang J. and Atkinson M., “Large-scale Comparative Sentiment Analysis of News Articles”

[17] Yan-Tak A., Jordan M., “On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes”

[18] https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/

[19] Platt J., Microsoft Research. Microsoft Way, Redmond , WA 98052 USA. http://www.research.microsoft.com/~jplatt

[20]https://turi.com/learn/userguide/supervised-learning/logisticregression.html

[21]https://turi.com/learn/userguide/supervisedlearning/boosted_trees_regression.html

[22]Campbell, C., Jiang R., Wright W.,Yan-Tak A, Ahres Y.,” Predicting Volatility in Equity Markets Using Macroeconomic News”.